Reconstruction of spatiotemporal capture data by means of orthogonal functions: the case of skipjack tuna (*Katsuwonus pelamis*) in the central-east Atlantic

UNAI GANZEDO 1, OIHANE ERDAIDE 2, AARON TRUJILLO-SANTANA 3, AIDA ALVERA-AZCÁRATE 4 and JOSÉ J. CASTRO 3

1 Department of Fisheries Management and Marine Research, Echebastar fleet s.l.u. Muelle Erroxape s/n (Box 39), 48370-Bermeo, Spain.
2 Research Center for Experimental Marine Biology and Biotechnology, Plentzia Marine Station (PIE), University of the Basque Country (UPV/EHU), Arratza, zg E-48620, Plentzia, Spain.
3 Faculty of Marine Sciences, University of Las Palmas de Gran Canaria, Edf. Ciencias Básicas, Campus Universitario de Tafira, 35017 Las Palmas de Gran Canaria, Spain. E-mail: jcastro@pesca.gi.ulpgc.es
4 Department of AGO-GHER, University of Liège, Allée du 6 Août 17, B5, Sart Tilman, 4000, Liège, Belgium.

SUMMARY: The information provided by the International Commission for the Conservation of Atlantic Tunas (ICCAT) on captures of skipjack tuna (*Katsuwonus pelamis*) in the central-east Atlantic has a number of limitations, such as gaps in the statistics for certain fleets and the level of spatiotemporal detail at which catches are reported. As a result, the quality of these data and their effectiveness for providing management advice is limited. In order to reconstruct missing spatiotemporal data of catches, the present study uses Data INterpolating Empirical Orthogonal Functions (DINEOF), a technique for missing data reconstruction, applied here for the first time to fisheries data. DINEOF is based on an Empirical Orthogonal Functions decomposition performed with a Lanczos method. DINEOF was tested with different amounts of missing data, intentionally removing values from 3.4% to 95.2% of data loss, and then compared with the same data set with no missing data. These validation analyses show that DINEOF is a reliable methodological approach of data reconstruction for the purposes of fishery management advice, even when the amount of missing data is very high.

Keywords: catches, missing data, spatiotemporal data reconstruction, singular value decomposition, DINEOF.

RESUMEN: La información proporcionada por la Comisión Internacional para la Conservación del Atún Atlántico (ICCAT) sobre capturas de atún bonito-listado (*Katsuwonus pelamis*) en el Atlántico Centro Oriental tiene una serie de limitaciones, tales como brechas en los estadísticas para ciertas flotas y el nivel de detalle espatio-temporal con el que se reportan las capturas. Como resultado, la calidad de estas datos y su efectividad para el asesoramiento en la gestión pesquera es limitada. Para reconstruir datos espatio-temporales perdidos de capturas, este estudio utiliza la técnica de interpolación por Funciones Ortogonales Empíricas (DINEOF), una técnica para la reconstrucción de datos perdidos, aplicada por primera vez a series de captura de pesca. DINEOF se basa en una descomposición en Funciones Ortogonales Empíricas (EOF) realizada con un método Lanczos. DINEOF fue probado con diferentes cantidades de datos perdidos, intencionalmente eliminando valores del 3.4% al 95.2% de la pérdida de datos, y luego comparado con el mismo conjunto de datos sin datos perdidos. Estas análisis de validación muestran que DINEOF es una metodología fiable para la reconstrucción de series de datos de origen pesquero, incluso cuando la cantidad de datos perdidos es muy alta.

Palabras clave: capturas, datos perdidos, reconstrucción de datos espacio-temporales, descomposición de valores singulares, DINEOF.
INTRODUCTION

In order to reduce the error range of their estimations, most population dynamics models (e.g. Virtual Population Analysis in Beverton and Holt 1957) require a big quantity of historical data of fishing effort and captures carried out by the different fishing fleets that target these stocks (Cushing 1983, Pitcher and Hart 1983, Hilborn and Walters 1992, Farrugio 1993, Lleonart 1993, Quinn and Deriso 1999, Cadima 2003). Unfortunately, there is not always enough historical data available to run these estimations and to effectively assess and manage the fisheries stock and its status (ICES 2005, Kelly and Codling 2006). There are a number of other potential problems associated with incomplete catch reports and/or biased reports, including the effect of regulation on catch reports (ICCAT 2008). Moreover, often the available data do not have suitable temporal continuity or the required quality (Scandol 2004, Daw and Gray 2005). The first Yearbook of World Captures edited by Food and Agriculture Organization (FAO) was published in 1950, as a partial solution to this problem. However, this and the following yearbooks of captures have many deficiencies due to the lack of data related to discards, sport fishing and unreported or unrecorded catches (Pauly 2009). For example, under-reports of anchovy catches given by Peru between 1951 and 1982 (Castillo and Mendo 1987), or those over-reported by China during the 1990s, as a consequence of its regional organization of fisheries statistical recording systems (Pang and Pauly 2001).

The majority of stock assessments are influenced by systematic errors due to a variety of causes, starting from the exactness of the capture data (Watson and Pauly 2001, Pauly 2009, Agnew et al. 2009), and from the discontinuities of the series. The most common source of uncertainty is found in the information systems, due to intentional loss of capture records, their manipulation or deliberate falsification, or just the fact that this information is inaccessible. Moreover, there is a lack of records on discards (Kennely and Broadhurst 2002, Kelleher 2005), recreational fishing captures (Morales-Nin et al. 2005, Lloret et al. 2008) and illegal catches that could exceed 40% of the total catch recorded in fishing areas such as northwest Africa (Agnew et al. 2009). Furthermore, although there are estimates of the relationship between landed and actual catch, the uncertainty in this relationship is very high (Patterson et al. 2001). Nevertheless, it is possible to calculate systematic errors (those due to the sampling methods) associated with the capture values during certain processes of stock assessment. Examples of this are the processes in which independent fisheries data are available (from scientific surveys), allowing the estimation of the associated uncertainty necessary for the forecasts (Anonymous 1999).

Several methods can be used to assess stocks in fisheries where there are very few data (Gómez-Muñoz 1990, Kelly and Codling 2006, Dowling et al. 2008) and where it is not possible to apply the complex analytical or probabilistic models in use (Lleonart 1993, Punt and Hilborn 1997, Daw and Gray 2005), which require a large quantity of detailed, exact and contrasted data. This often occurs in small-scale, exploratory, developing or expanding fisheries, where estimates of biomass are not available. However, neither of these methods can be used to make future stock assessments in order to establish a more realistic management strategy. According to Chen et al. (2003), deficiency in data quantity, and also quality, tends to yield biased assessments of the status of fisheries stocks and increase the uncertainty in stock assessment, subsequently complicating the identification of a more suitable management strategy. The data quality is affected not only by the truthfulness of the information sources but also by their temporal continuity. For example, the use of covariates as a method to provide additional information about model parameters (i.e. environmental effects) to complement the fisheries stock assessment models is strongly conditioned by missing values in the starting series (Maunder and Deriso 2010). Maunder and Deriso (2010), after testing different alternatives to deal with missing values, concluded that the random effects method does not provide substantial benefit over other less complex methods such as ignoring the years with missing covariate values, substituting the missing values with the mean of the observed ones, and estimating the missing values as free parameters. However, the same authors also pointed out that if the missing data are substituted with inappropriate values, the relationship with the covariates may be distorted, and the uncertainty may be underestimated (Beckers and Rixen 2003). Therefore, it is important to develop appropriate methods to deal with missing values.

The main objective of this study is to validate the Data Interpolating Empirical Orthogonal Functions (DINEOF) method of missing data reconstruction, applied here for first time to fisheries data, as an effective method to compensate the data loss observed in the historical catch series (Alvera-Azcárate et al. 2007). To achieve this objective the method was applied to the catch series of skipjack tuna (Katsuwonus pelamis) from the central-east Atlantic provided by the International Commission for the Conservation of Atlantic Tunas (ICCAT). ICCAT’s information on catch has a number of limitations, most commonly related to the completeness of statistics for certain fleets, the level of spatiotemporal detail available and the taxonomic level at which catches are reported. In order to improve these data and their usefulness in providing management advice, there is a need to develop reliable methods for interpolation and extrapolation of the available data.
MATERIALS AND METHODS

Data Set

Catch data of skipjack tuna (*Katsuwonus pelamis*) reported in the FAO fishing area 34 and the Azores (Portugal, Central-east Atlantic) were selected for this study. These data were obtained from the free access database of ICCAT. Seasonal skipjack captures from 1980 to 2006 were reported in an area ranging from 35°N to 5°S, and from 40°W to 10°E, with a resolution of 5° × 5° (Fig. 1). There were 16 time series (season from spring 1980 to autumn 2006) for each quadrant (5° × 5°) available. These 16 time series have at least 90% of good data (without missing values). This data set has 16 spatial series with 107 time values each (the initial matrix was 16 × 107 = 1712 data). The aim was to reconstruct the complete data set (1712 values) that has 58 missing values (3.4%).

Spatiotemporal catch data of skipjack tuna were used to evaluate the behaviour of the DINEOF using real data and randomly removing different amounts of them. The removed values were compared with the estimated (or reconstructed) ones from the DINEOF analysis. Finally, these data were used to study the behaviour of the model for capture series from two regional fisheries in the central-east Atlantic: off the Canary Islands (27.5°N 17.5°W) and the Gulf of Guinea (2.5°N 2.5°E) (the areas with thick black line circles in Fig. 1).

DINEOF application

DINEOF is a self-consistent method for reconstructing missing values in oceanographic data sets (Beckers and Rixen 2003), based on the fact that an optimal number of Empirical Orthogonal Functions (EOFs), usually very small if compared to the total number of EOFs, retains a large fraction of the total variance of the whole data set. The information contained in the data set is used by the EOF series to infer the missing values, reducing the possibility of biasing the time series of catch data and increasing the representativeness of the data set. This method is already implemented in Fortran and freely available (http://modb.oece.ulg.ac.be/mediawiki/index.php/DINEOF), and the advantages of using a Lanczos eigensolver have already been shown by Alvera-Azcárate et al. (2005).

The DINEOF method fills the missing data by means of an iterative process. Initially, the leading EOF is computed after substituting missing values by zeros. Then, the missing data are substituted by leading EOF values at the corresponding locations. The process is iterated until convergence in the anomalies at the missing values is achieved from one iteration to the next within a prescribed tolerance level. Once the convergence is reached, the number of computed EOFs increases from 1 to 2 and so on up to \( k_{\text{max}} \) EOFs. At the end, there is an estimate for the missing data reconstructed after convergence is achieved with a reconstruction computed using \( 1, 2, \ldots, k_{\text{max}} \) EOFs. The optimum number of EOFs to be used in the reconstructions is defined by means of the cross-validation technique (Wilks 1995). In this case, with 3.4% of missing data, the optimum number of EOFs was four.

For this study, 3% of good data are set aside from the reduced data set to be compared later with the reconstructed values (sensus Alvera-Azcárate et al. 2005). The optimal number of EOFs is the one that minimizes the error between the data set maintained aside and the values obtained at these points with the reconstruction method. Once the optimal number \( k_{\text{max}} \) of EOFs is known, the whole procedure is repeated again, this time including the data set maintained aside for cross-validation, but using only the \( k_{\text{max}} \) leading EOFs considered as optimal. Final values for the missing data are then computed. In this application, the EOF decomposition is performed by means of a Lanczos iterative eigensolver (Toumazou and Cretaux 2001), which allows optimal CPU use for the computation of EOFs in large matrices, since only the leading singular vectors must be computed. This method was successfully applied to sea surface temperature images of the Adriatic by Alvera-Azcárate et al. (2005).

We decided to work with the logarithm of the likelihood function, known as log-likelihood. Since the logarithm is a strictly increasing function, the same parameter values will maximize both the likelihood and log-likelihood functions.

Scaling of EOFs for graphical representation

EOFs (spatial loading factors) are obtained from the singular value decomposition of the data covariance...
matrix as a set of orthonormal column vectors:

\[
\hat{e}_i(r_j) = \begin{pmatrix} e_{1j} \\ \vdots \\ e_{Mj} \end{pmatrix}
\]  

(1)

where \(e_{ij}\) is the value of the \(i^{th}\) empirical orthogonal function at the \(j^{th}\) grid point, \(r_j\) represents one-dimensional all spatial points of the eigenvector, and \(M\) represents the numbers of grid points. To help in the interpretation of the results, the EOFs are scaled in two ways: each EOF is shown scaled by the units of the field it is representing (metric tons in our case) by means of the expression:

\[
\sqrt{\lambda_i} \hat{e}_i
\]

(2)

which is equivalent to expressing the EOF as the regression of the corresponding principal component (PC) (used in Fig. 2) onto the original catches field. In this equation, \(\lambda_i\) is the eigenvalue of the covariance matrix associated with the \(i^{th}\) eigenvector \(\hat{e}_i\). This scaling (used in Fig. 3) allows an easy visual detection of the points where each EOF involves a substantial effect on the reconstructed data.

**Evaluation of reconstruction skill**

Finally, in order to obtain a measure of the accuracy of the reconstructed fields, two parameters were used as a measure of the error. The first one is the root mean square error (RMSE). As the RMSE can be very sensitive to outliers in the sample, the mean absolute deviation (MAD) was also used. Those two error estimates were computed between the original values and the reconstructed estimates (cross-validation). These forms of expression of the error have the advantage that they retain the units of the variable and are thus easily interpretable as typical error magnitude. Also, the comparison was quantified by means of the Pearson correlation coefficient (\(r\)) and the \(p\)-value (Wilks 1995).

**RESULTS**

The variance explained by the reconstruction, using real data with 3.4% of missing values, was 76.6% with 4 EOFs retained.

<table>
<thead>
<tr>
<th>EOFs</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>% EOFs</td>
<td>38.26</td>
<td>21.15</td>
<td>11.86</td>
<td>5.32</td>
<td></td>
<td></td>
<td>76.59</td>
</tr>
</tbody>
</table>

**Table 1.** Variance explained for each EOF approximation. Using EOF with four EOFs retained.
four EOFs retained (Table 1). The first and second PCs (PC1 and PC2, Fig. 2 A,B) emphasize the seasonality of the fishery, a consequence of the migratory pattern of the skipjack from the Gulf of Guinea to water north of the Azores (northeast Atlantic) as summer approaches (González-Ramos 1992, EOF1 and EOF2,
In winter the highest catches were recorded in the southern part of its geographical distribution (Gulf of Guinea) and in spring-summer in the northern part (Canary Islands-Azores-Cantabrian Sea) (Gouveia and Mejuto 2003).

Measured in terms of fraction of reconstructed variance, the contribution of EOF1 (38.26% of total variance) was particularly high in Canary Island waters (Fig. 3A). EOF2 (Fig. 3B) was also important (21.15% of total variance) and explained fractions of variance in waters off Senegal and Gulf of Guinea not explained by EOF1. The contribution of EOF3 (Fig. 3C) (11.86% of total variance) was high off Senegal, while that of EOF4 (5.32% of total variance) was higher in the Gulf of Guinea (Fig. 3D).

### Sensitivity of the results to the missing values

The sensitivity of the reconstruction was more evident with the increase in missing data (1% of values were removed from each reconstruction, from 3.4% to 95.2%) (Table 2). The variance explained by the EOFs was relatively high, even when the percentage of missing data was close to 40% (Table 2 and Fig. 4A). Moreover, the number of EOFs retained ranged between 2 and 4, but at higher levels of missing data they were not very stable (Fig. 4B).

The validation was carried out on a data set with increasing amounts of missing values. The RMSE and MAD, as derived by cross-validation, increased as the number of missing values increased (Fig. 5). Above 60% of missing values in the database, the explained variance decreased notably (Fig. 4A). Correlations between real and EOF- reconstructed data series were higher than those obtained from the EOF approach, particularly when missing data were higher than 40%-50% (Fig. 6).

As a last step in the verification of the reconstruction, the behaviour of the scatter plot of the relationship between reconstructed (with 4 EOFs and 76.6% of explained variance) and observed data indicated that...
as the percentages of missing values increased the fit was worse (Fig. 7). The reconstructed values tended to be higher than the real ones. For most of the remaining data, including high catch values, the agreement between original and reconstructed data was good. Therefore, the correlation between original and reconstructed data in the Canary Islands area was better than in the Gulf of Guinea (Figs 8 and 9).

**DISCUSSION**

The results shown in this study indicate that DI-NEOF is a suitable method for reconstructing missing values in spatiotemporal fishing data series. The method is robust and simple to use, the code is freely available and it does not need any a priori information about the statistical error of the data. Also, this method can be used to analyse the spatiotemporal performance of a fishery, as we have shown for the skipjack tuna in the central-east Atlantic.

In the case of skipjack, the EOF principal components highlight the seasonality of the different fisheries analysed as a consequence of the periodic

![Fig. 7. Comparison of reconstructed data with original data: A, 3.39% of missing data; B, 17.87% of missing data; C, 32.36% of missing data; and D, 51.69% of missing data. The diagonal line represents the fit corresponding to a linear model (line y=x).](image)

![Fig. 8. Correlation between original and reconstructed data in the Canary Islands and the Gulf of Guinea.](image)
north-south migratory pattern of this species (as an oscillatory wave) along the central-east Atlantic, from the Gulf of Guinea to the Azores and the Cantabrian Sea, with the progressive northward water warming in spring and summer, and the journey back in autumn and winter (Cayré and Farrugio 1986, Cayré et al. 1986, Bard and Antoine 1986, Miyabe and Bard 1986, González-Ramos 1992, Gouveia and Mejuto 2003, Trujillo-Santana 2010). However, the model appears to be sensitive to the nature of the fishery, as data from the live bait fishery of the Canary Islands were rebuilt better than the industrial data developed in the Gulf of Guinea. This could be because the Canarian fishing fleet has remained almost unchanged over the study period, maintaining almost constant the unit of fishing effort (González-Ramos 1992, Pallarés et al. 2005, Trujillo-Santana 2010), which must necessarily be reflected in the temporal trends in catches and catch per unit effort (CPUE). However, the fishing fleets operating in the Gulf of Guinea show significant differences between them, in addition to different regulations (Miyake et al. 2004) that have affected fleets unevenly and could have increased uncertainty and thus the variance in the capture and indices of abundance over time. Moreover, in the Gulf of Guinea there are complex processes of reproduction, recruitment and behavioural interaction with the environment and fishing strategy that also could increase the variability of skipjack captures (Fonteneau and Pallares-Soubrier 1996, Solari et al. 2003). Ménard et al. (2000) suggested that the FAD fishery may have wide-ranging effects on the migration of tunas in general and on the productivity of the skipjack population in particular.
The behaviour of the scatter plot of the relationship between reconstructed and observed data suggests that for a high percentage of missing data, over 40%, the fit became worse, probably because the central limit theorem is not satisfied. The use of logarithms to normalize the series appears to be insufficient, so it might be necessary to use other techniques. Also, it has to be taken into account that these results could be different using other extended EOF approximations or different lags of the data, or adding CPUE data of the same species or other related ones. The approach leading to the best results depends on the data set used, so tests are needed to determine the best approach when using DINEOF. The use of extended EOFs (Alvera-Azcárate et al. 2007) can help in the development of a complete data set. And in this case, the use of extended EOFs on catch series with a time lag of one season has improved the results. It is foreseeable that the best results could be obtained using different variables along with the catch data (Alvera-Azcárate et al. 2007). Examples of multivariate approaches could include the use of CPUE, meteorological conditions, or even large-scale climate indicators such as the NAO (North Atlantic Oscillation) index. Future studies could test the possibility of improving the analysis by replacing the missing values with a linear interpolation in time or space with nearby cells. Comparing the performance of DINEOF with other simple mapping techniques such as Kriging is also surely an interesting topic for a future study.

ACKNOWLEDGEMENTS

Unai Ganzedo thanks the ECHEBASTAR FLEET company for providing financial support. Oihane Erdaide is the holder of a Zabalduz-2012 Fellowship (Resol. n° 3625/2012) granted by the University of the Basque Country (UPV/EHU). Help by AZTI’s tuna group (Marine Research Division) is also acknowledged, particularly to Haritz Arrizabalaga and Josu Santiago. Research by Universidad de Las Palmas de Gran Canaria staff was supported by Canary Islands CIE: Tricontinental Atlantic Campus.

Aaron Trujillo-Santana died in October 2010, and this work was expected to be part of his PhD thesis, but unfortunately he was unable to finish it. He received a grant from the Cabildo Insular of Gran Canaria, Spain.

We also want to thank an anonymous referee for valuable comments that allowed us to considerably improve this paper.

REFERENCES


Chen Y., Chen L., Stergiou K.I. 2003. Impact of data quantity on indicators such as the NAO (North Atlantic Oscillation) index. Future studies could test the possibility of improving the analysis by replacing the missing values with a linear interpolation in time or space with nearby cells. Comparing the performance of DINEOF with other simple mapping techniques such as Kriging is also surely an interesting topic for a future study.

ACKNOWLEDGEMENTS

Unai Ganzedo thanks the ECHEBASTAR FLEET company for providing financial support. Oihane Erdaide is the holder of a Zabalduz-2012 Fellowship (Resol. n° 3625/2012) granted by the University of the Basque Country (UPV/EHU). Help by AZTI’s tuna group (Marine Research Division) is also acknowledged, particularly to Haritz Arrizabalaga and Josu Santiago. Research by Universidad de Las Palmas de Gran Canaria staff was supported by Canary Islands CIE: Tricontinental Atlantic Campus.

Aaron Trujillo-Santana died in October 2010, and this work was expected to be part of his PhD thesis, but unfortunately he was unable to finish it. He received a grant from the Cabildo Insular of Gran Canaria, Spain.

We also want to thank an anonymous referee for valuable comments that allowed us to considerably improve this paper.

REFERENCES


Chen Y., Chen L., Stergiou K.I. 2003. Impact of data quantity on indicators such as the NAO (North Atlantic Oscillation) index. Future studies could test the possibility of improving the analysis by replacing the missing values with a linear interpolation in time or space with nearby cells. Comparing the performance of DINEOF with other simple mapping techniques such as Kriging is also surely an interesting topic for a future study.

ACKNOWLEDGEMENTS

Unai Ganzedo thanks the ECHEBASTAR FLEET company for providing financial support. Oihane Erdaide is the holder of a Zabalduz-2012 Fellowship (Resol. n° 3625/2012) granted by the University of the Basque Country (UPV/EHU). Help by AZTI’s tuna group (Marine Research Division) is also acknowledged, particularly to Haritz Arrizabalaga and Josu Santiago. Research by Universidad de Las Palmas de Gran Canaria staff was supported by Canary Islands CIE: Tricontinental Atlantic Campus.

Aaron Trujillo-Santana died in October 2010, and this work was expected to be part of his PhD thesis, but unfortunately he was unable to finish it. He received a grant from the Cabildo Insular of Gran Canaria, Spain.

We also want to thank an anonymous referee for valuable comments that allowed us to considerably improve this paper.

REFERENCES


